Final Project: Brain MRI segmentation

Shlomo Yehezkel
Department of Computer Science
Reichman University
shlomo.yehezkel@post.runi.ac.il

Oz Sharlin
Department of Computer Science
Reichman University
oz.sharlin@post.runi.ac.il

Nadav Loebl
Head of AI
Beilinson Innovation
nadavloe@clalit.org.il

Abstract

Brain tumor segmentation is essential for the diagnosis and prognosis of patients with gliomas. Among MRI modalities, FLAIR MRI is widely available and routinely used in most brain MRI scans, while pre-contrast and post-contrast MRI are also frequently utilized in clinical practice. However, post-contrast MRI requires the injection of a gadolinium-based contrast agent, which presents certain clinical disadvantages and may pose health risks for patients with kidney failure.

In this study, we developed a segmentation method based on FLAIR-only MRI that outperforms multi-modal MRI models that incorporate post-contrast imaging, achieving a Dice score improvement of nearly 3% on average. Our approach leverages 3D contextual information and combines ensemble techniques to enhance tumor segmentation accuracy for FLAIR-only models. Specifically, we introduce a novel segmentation strategy that captures volumetric information by processing MRI scans in chunk-based segments, where each chunk consists of the target slice and its neighboring sequential slices. This method ensures adaptability to varying scan resolutions, improving segmentation robustness. Additionally, our method significantly outperforms the baseline multi-modal MRI model [1] by nearly 9%.

The proposed FLAIR-only MRI approach offers significant clinical advantages, including simplifying the scanning procedure, reducing scan duration, and eliminating the need for contrast agent injection, thereby minimizing patient discomfort and reducing potential health risks.

NeurIPS 2021 AI for Science Workshop.

1 Introduction

1.1 MRI scan for brain tumor segmentation

Brain tumors are a significant health concern worldwide, affecting thousands of individuals each year. According to epidemiological studies, primary brain tumors occur in around 250,000 people a year globally and make up less than 2% of cancers [21]. The prevalence of brain tumors varies based on factors such as age, genetics, and environmental influences. Primary brain tumors, which originate within the brain, are classified into benign and malignant types. Among the malignant tumors, gliomas—including glioblastoma multiforme (GBM)—are the most aggressive and challenging to treat. Other types of brain tumors include meningiomas, astrocytomas, and pituitary adenomas, each presenting distinct clinical challenges and treatment considerations.

1.2 The Impact of Brain Tumors on Health and Treatment Approaches

The implications of brain tumors are profound, affecting cognitive function, motor skills, and overall quality of life. Symptoms can range from persistent headaches and seizures to changes in personality and memory loss. The severity of these symptoms often depends on the tumor's location, size, and growth rate. Early detection is crucial for improving prognosis and guiding treatment options, which may include surgery, radiation therapy, and chemotherapy.

Magnetic Resonance Imaging (MRI) plays a pivotal role in the detection and diagnosis of brain tumors. MRI provides high-resolution images of brain structures, allowing clinicians to visualize tumors with exceptional detail. Compared to other imaging modalities like CT scans, MRI offers superior contrast differentiation between normal and abnormal tissues, making it the preferred choice for brain tumor assessment. MRI is widely used in clinical practice due to its non-invasive nature and ability to capture multiple imaging sequences, enhancing diagnostic accuracy.

1.3 Types of MRI scans for Tumor Diagnosis

MRI scans for brain tumor detection typically include three key sequences: pre-contrast (Pre), fluid-attenuated inversion recovery (FLAIR), and post-contrast (Post). The Pre scan provides a baseline image before any contrast agent is administered, capturing structural details of the brain. FLAIR imaging enhances the visibility of lesions by suppressing cerebrospinal fluid signals, making abnormalities such as edema and non-enhancing tumors more distinguishable. The Post scan, performed after administering a contrast agent, highlights regions with abnormal blood-brain barrier permeability, helping to differentiate tumor tissue from surrounding structures. The combination of these sequences enables comprehensive tumor assessment, aiding in accurate diagnosis and treatment planning.

1.4 Contrast Agent Injection

The use of contrast agents in MRI, such as Gadolinium-Based Contrast Agents (GBCAs), enhances tumor visibility and differentiation from healthy tissue. However, the injection process has implications for patients, including time, convenience, and potential side effects.

The administration of contrast agents adds additional time to the MRI procedure, usually extending the scan by 10–15 minutes. This includes preparation, injection, and post-injection imaging.

While GBCAs are generally safe, some patients may experience mild side effects such as nausea or dizziness. Rarely, allergic reactions or nephrogenic systemic fibrosis (in patients with kidney disease) may occur.

MRI scans produce images in multiple planes to offer comprehensive visualization of brain structures. There are three main intersectional planes.

1.5 MRI Planes

Axial Plane: Also known as the transverse plane, this horizontal view slices the brain from top to bottom, providing a detailed cross-sectional view of tumor positioning.
Sagittal Plane: This vertical plane divides the brain into left and right halves, allowing for a lateral perspective crucial for evaluating midline structures.
Coronal Plane: A frontal view that divides the brain into front and back sections, offering insights into symmetrical tumor growth and ventricular system abnormalities.

Each of these planes plays a crucial role in tumor assessment, ensuring precise localization and aiding in surgical and therapeutic planning.

1.6 MRI Resolution

MRI resolution is a critical factor in detecting and assessing brain tumors. The scanning resolution is determined by parameters such as voxel size, slice thickness, and field of view (FOV), which impact image clarity and diagnostic accuracy.

High resolution magnetic resonance imaging (1 mm or smaller voxel size): Provides exceptional detail, enabling precise tumor boundary delineation. It is preferred for pre-surgical planning and research applications.
Standard Clinical Resolution (1–3 mm voxel size): Balances scan time with diagnostic quality, commonly used in routine clinical practice.
Low-Resolution MRI (>3 mm voxel size): Faster scanning but lower detail, typically used in emergency settings where rapid assessment is required.

1.7 The importance of automatic and accurate MRI segmentation

One of the critical challenges in brain tumor analysis is accurate segmentation, which involves delineating tumor boundaries within MRI scans. Automated segmentation techniques are essential for improving diagnostic efficiency, facilitating treatment planning, and monitoring tumor progression. Manual segmentation by radiologists is time-consuming and prone to variability, necessitating the development of advanced computational methods. Deep learning and machine learning algorithms have emerged as powerful tools for automating tumor segmentation, significantly enhancing precision and reproducibility.

The need for effective brain tumor segmentation is further underscored by its impact on personalized medicine. Precise segmentation enables clinicians to tailor treatments to individual patients, optimizing therapeutic strategies while minimizing damage to surrounding healthy tissues. As research advances, the integration of artificial intelligence in MRI analysis continues to revolutionize the field, offering promising avenues for early detection and improved patient outcomes.

1.8 Neurologist clinical insights on the use of MRI for brain tumor segmentation

To gain a deeper understanding of the Brain MRI scanning procedure, we consulted a neurologist from Beilinson Hospital, Israel. According to the neurologist, FLAIR MRI is widely available and routinely used in most brain MRI scans, while pre-contrast and post-contrast MRI are also frequently utilized in clinical practice. However, post-contrast MRI requires the injection of a gadolinium-based contrast agent, which, according to recent evidence, may accumulate in brain tissue, though its toxic level is still unclear. Additionally, gadolinium-based contrast agents are not recommended for patients with kidney failure due to potential health risks. Specifically, for Lower Grade Gliomas, post-contrast and FLAIR MRI are essential for diagnosis.

We further inquired whether there might be clinical advantages to using FLAIR-only MRI. The neurologist highlighted several benefits, including simplifying the scanning procedure, reducing scan duration, and eliminating the need for contrast agent injection, thereby minimizing patient discomfort and potential health risks.

2 Objectives

In this project, we aim to facilitate medical procedures by optimizing segmentation performance using only FLAIR MRI, eliminating the need for post-contrast MRI, which requires the injection of a contrast agent. As previously mentioned, this could offer significant clinical advantages.

We explore brain MRI segmentation techniques with a focus on utilizing 3D contextual information to improve the performance of FLAIR-only models. Our goal is to develop methods that can achieve segmentation results comparable to those obtained with multi-modal MRI, which includes post-contrast imaging. By leveraging spatial relationships within volumetric data, we seek to enhance tumor detection and classification, ultimately contributing to improved diagnostic accuracy and treatment planning for brain tumor patients.

3 Related Works

Semantic segmentation in images is a well-established task in computer vision with numerous applications in medical imaging. Over the years, deep learning-based approaches have significantly advanced segmentation performance, particularly in 2D medical image segmentation using convolutional neural networks (CNNs).

3.1 Deep Learning-Based Segmentation

One of the most impactful architectures for biomedical image segmentation is U-Net, introduced in 2015 [18]. U-Net follows a U-shaped encoder-decoder structure with skip connections that help retain spatial information lost during downsampling. Since its introduction, U-Net has become the foundation for many medical imaging applications, particularly for MRI and CT segmentation tasks.

In recent years, transformer-based architectures, originally developed for natural language processing [17], have been successfully adapted to computer vision. The Vision Transformer (ViT) [15] introduced self-attention mechanisms for image processing, treating images as sequences of patches. Several transformer-based architectures have been developed for semantic segmentation, including SegFormer (2021) [16], which integrates a hierarchical transformer encoder with a lightweight decoder to efficiently fuse multi-scale features.

3.2 Foundation Models for Segmentation

The emergence of foundation models has further transformed the field of image segmentation. Segment Anything [14], introduced by Meta in 2023, is a transformer-based segmentation model trained on a large-scale dataset to enable prompt-based segmentation across diverse domains. Following this, MedSAM (Segment Anything for Medical Images) [12] was proposed in 2023 as a foundation model for medical image segmentation. MedSAM was trained on a dataset containing over one million image-mask pairs spanning various modalities, such as CT, MRI, and ultrasound, enabling zero-shot generalization to new tasks. An improved version, MedSAM2 [13], was later introduced, enhancing zero-shot segmentation capabilities by incorporating video-based auto-tracking mechanisms for both 2D and 3D medical images.

3.3 Brain Tumor Segmentation and the BRATS Challenge

Brain tumor segmentation is a critical task in medical imaging, commonly evaluated using the Brain Tumor Segmentation (BRATS) dataset [11], which has been a core part of the MICCAI Brain Tumor Segmentation Challenge since 2012. The dataset provides multimodal MRI scans (T1, T1c, T2, and FLAIR) with corresponding ground-truth annotations for tumor regions. Annual updates for this dataset incorporate more annotated cases. Widely used versions include BRATS 2020 and BRATS 2021,

Several deep learning-based approaches have been developed for BRATS segmentation. A dominant framework in recent years is nnU-Net[5] , which automatically adapts its architecture, preprocessing, and post-processing to different datasets without requiring extensive manual tuning. The winning model of the BRATS 2020 challenge was based on nnU-Net [4], utilizing a 3D U-Net architecture with 3D convolutions instead of traditional 2D convolutions.

Subsequent improvements were introduced in the BRATS 2021 challenge, where group normalization replaced batch normalization, and axial attention mechanisms were added to the decoder [2]. The BRATS 2022 challenge saw the introduction of an ensemble [6] approach combining DeepSeg [3], nnU-Net, and DeepSCAN, which achieved state-of-the-art performance on two unseen test datasets.

3.4 Generative AI for Medical Image Segmentation

With the rise of generative AI, techniques such as Generative Adversarial Networks (GANs) and diffusion models have been explored to address the limited availability of annotated medical data. Generative models allow for synthetic data augmentation, improving model robustness and generalization.

A notable approach in BRATS 2021 improved 3D U-Net segmentation performance by generating adversarial examples [7] using a GAN-based augmentation strategy. A submission to BRATS 2023 [8]introduced a GAN-based synthetic augmentation method conditioned on segmentation labels, combined with an ensemble of state-of-the-art models. Other studies have explored the use of StyleGAN (versions 1–3) and diffusion models for synthetic MRI scan generation [9], demonstrating competitive segmentation performance on BRATS 2020 and BRATS 2021 datasets.

3.5 LGG Segmentation Dataset

Another open dataset for brain tumor segmentation is the Lower-Grade Glioma (LGG) Segmentation Dataset [10], which includes brain MRI scans with manual FLAIR abnormality segmentation masks. The dataset also provides patient metadata and genomic information.

A 2019 study [1] demonstrated that deep learning-extracted MRI features could predict tumor molecular subtypes in lower-grade gliomas, showcasing the potential for non-invasive imaging-based genomic analysis in brain cancer. The proposed model used a 2D U-Net architecture, incorporating preprocessing techniques such as skull stripping (removal of the skull) and post-processing with connected components analysis to remove false positives. Due to the dataset's limited size, a cross-validation strategy was applied to ensure statistical robustness.

In this paper, we utilize the LGG dataset to explore the benefits of 3D contextual information in tumor segmentation. Additionally, we investigate whether accurate segmentation can be achieved without post-injection MRI scans, which could have significant clinical implications by reducing the need for contrast-enhanced imaging in certain diagnostic workflows.

4 Dataset

4.1 Data Acquisition

The LGG Segmentation dataset consists of brain magnetic resonance (MR) images along with manually annotated segmentation masks for FLAIR abnormalities. The data was obtained from The Cancer Genome Atlas (TCGA) lower-grade glioma collection. It includes MR images from 110 patients who had at least one available fluid-attenuated inversion recovery (FLAIR) sequence and corresponding genomic cluster data.

4.2 Data Format and Organization

All images are provided in `.tif` format, with each image containing three channels. The dataset comprises a total of 3,929 images. The three channels in each image correspond to different MR sequences:

- **Pre-contrast**
- **FLAIR**
- **Post-contrast**

For 101 cases, all three sequences are available. However, for 9 cases, the post-contrast sequence is missing, and for 6 cases, the pre-contrast sequence is missing. To maintain consistency across all images, missing sequences were replaced with the FLAIR sequence, ensuring that each image retains three channels.

4.3 Segmentation Masks

The dataset also includes binary segmentation masks, provided as single-channel images, that highlight FLAIR abnormalities in the FLAIR sequence. These manually annotated masks serve as ground truth labels for training and evaluating segmentation models.

4.4 Folder Structure and Naming Conventions

The dataset is organized into 110 folders, each named according to the corresponding case ID, which also contains information about the source institution. Each folder contains the MR images following a standardized naming convention. Additionally, a `data.csv` file is provided, containing patient data and tumor genomic cluster information.

This dataset serves as a valuable resource for developing and evaluating automated segmentation models for brain tumor detection in MR images.

4.5 Visualizations

The following image presents a visualization of brain MRI segmentation for tumor detection. The first row displays the original MRI scans, followed by three key imaging modalities:

- **Pre-Contrast:** Shows the brain's structure before contrast enhancement.
- **FLAIR:** Highlights fluid-attenuated regions to enhance abnormal tissue visibility.
- **Post-Contrast:** Accentuates tumor-affected areas using contrast agents.

The following image consists of five rows, each representing a different type of MRI scan or its corresponding segmentation. The first row displays the raw MRI scan, which serves as the unprocessed input data. The second row corresponds to the pre-contrast MRI, providing anatomical details before the administration of a contrast agent. The third row represents the FLAIR (Fluid-Attenuated Inversion Recovery) sequence, which enhances lesion visibility by suppressing fluid signals, making it particularly useful for detecting abnormalities such as white matter lesions. The fourth row shows the post-contrast MRI, which highlights regions with contrast uptake, often aiding in the identification of pathological structures like tumors. Finally, the fifth row contains the segmentation mask, which delineates the target regions, facilitating automated analysis and interpretation of brain structures.

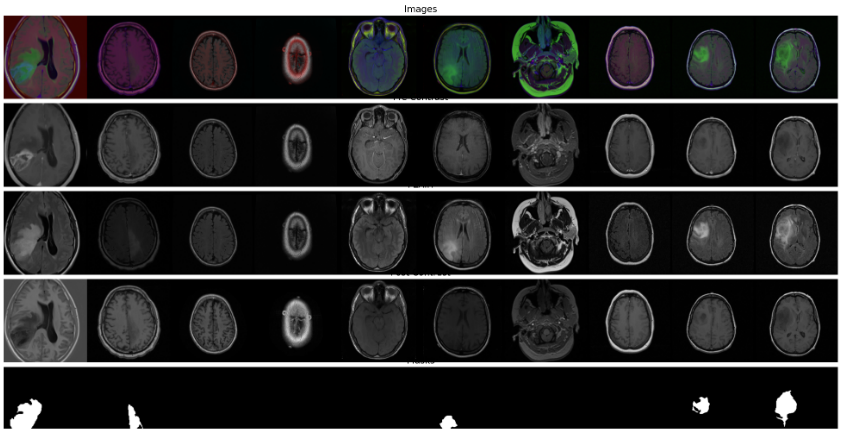

Figure 1: Brain MRI Segmentation: Raw Images, Pre, FLAIR, Post and mask for Tumor Detection

To illustrate the concept of brain slices, we generated a 3D model of the head by stacking 2D MRI slices from different planes, creating a volumetric representation of the head

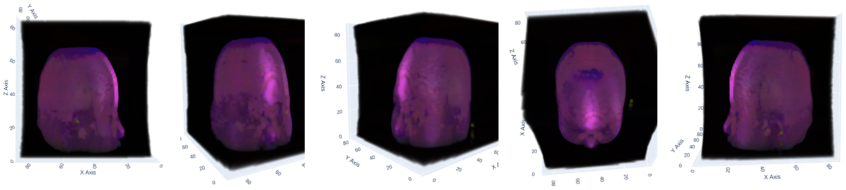

Figure 2: Volumetric Reconstruction of the Head from 2D MRI Slices

5 Method

The method described in this section is applied to two types of models:

1. Models that process only FLAIR MRI samples (1-channel per MRI slice).
2. Models that process pre-contrast + FLAIR + post-contrast samples (3-channel slices).

5.1 3D Method

In this work, we aim to leverage 3D contextual information for tumor segmentation prediction. Our goal is to design a solution that is both robust and adaptable to the varying MRI scanning resolutions in our dataset, where scans range from 20 to 88 slices per patient.

5.1.1 Approach Considerations

One approach we considered was standardizing the number of MRI slices per patient, enabling the use of a fully 3D model that processes stacked slices as a single 3D input to predict an entire 3D segmentation mask. This approach has been widely used in studies based on the BRATS dataset, which contains a fixed number of slices per scan. However, due to the variable number of slices in our dataset, this method presented challenges:

- Padding to 88 slices (the maximum in our dataset) would be inefficient, leading to unnecessary computation on non-informative slices.
- Sampling or unifying layers to 20 slices (the minimum) could result in significant loss of critical information.

5.1.2 Proposed 3D Chunk-Based Approach

Instead of forcing uniformity in slice count, we propose a "chunk-based" approach that predicts the tumor segmentation mask for each MRI slice while incorporating 3D spatial context.

Each input chunk consists of:

- The target slice (the slice for which the segmentation mask is predicted).
- Neighboring slices, symmetrically stacked above and below the target slice.

The chunk size is always an odd number to maintain symmetry around the target slice. This approach ensures that models receive 3D context, while remaining adaptable to varying scan resolutions without requiring preprocessing transformations.

5.1.3 Chunk Size Exploration

We experiment with various chunk sizes, ranging from 1 slice to 19 slices, incrementing by 2 slices at each step:

- Chunk size of 1 slice represents a 2D input with no 3D context.
- Chunk size of 19 slices is the maximum tested, as some patient scans contain only 20 slices.

5.1.4 Inference Strategy

During inference, patient scans are processed chunk-by-chunk, shifting by one slice at a time. The model is applied to each chunk, and the predicted segmentation masks for the target slices are aggregated to generate the final tumor segmentation for the entire MRI scan.

This approach is applicable only to internal slices, as edge slices lack sufficient contextual information. The required context depends on the chunk size used by the model.

5.2 Model Architecture

During implementation, we considered various architectural options. While transformers have demonstrated strong performance in vision tasks, we ultimately selected UNet due to practical constraints, including a limited dataset size and hardware specifications. Model training was conducted using Google Colab with an Nvidia A100 GPU (40 GB RAM). Training took approximately 25 minutes for 100 epochs, with some variation depending on the configuration.

We employ both 2D and 3D U-Net architectures, applied to 2D and 3D inputs, respectively.

- For chunk size = 1 slice → 2D U-Net is used.
- For chunk size > 1 slice → 3D U-Net is used.

Both 2D and 3D U-Net variants share a similar architecture, consisting of:

- 3 encoder blocks and 3 decoder blocks,
- Downsampling by a factor of 8 in the encoder,
- Upsampling back to the original size in the decoder,
- Final sigmoid activation to predict the foreground probability.

The key difference between the models is:

- 2D U-Net → Uses 2D convolutional kernels.
- 3D U-Net → Uses 3D convolutional kernels to capture volumetric information.

A diagram illustrating the architecture is provided below:

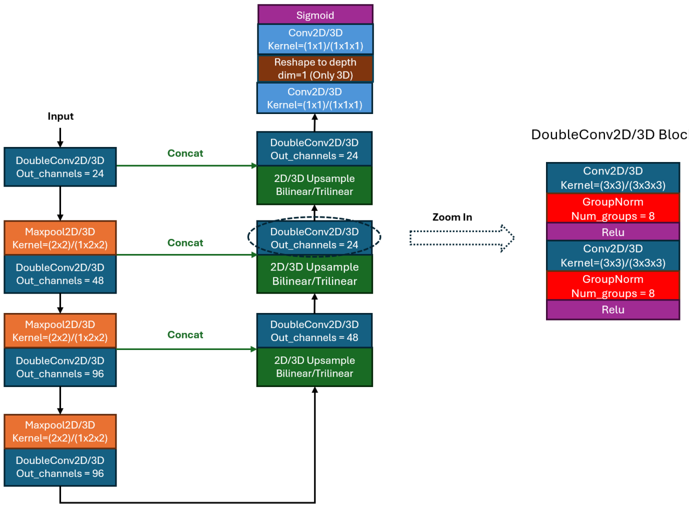
Sigmoid
Conv2D/3D Kernel=(1x1)/(1x1x1)
Reshape to depth dim=1 (Only 3D)
Conv2D/3D Kernel=(1x1)/(1x1x1)
Input
DoubleConv2D/3D Out_channels = 24
Concat
DoubleConv2D/3D Out_channels = 24
2D/3D Upsample Bilinear/Trilinear
Maxpool2D/3D Kernel=(2x2)/(1x2x2)
DoubleConv2D/3D Out_channels = 48
Concat
DoubleConv2D/3D Out_channels = 24
2D/3D Upsample Bilinear/Trilinear
Maxpool2D/3D Kernel=(2x2)/(1x2x2)
DoubleConv2D/3D Out_channels = 96
Concat
DoubleConv2D/3D Out_channels = 48
2D/3D Upsample Bilinear/Trilinear
Maxpool2D/3D Kernel=(2x2)/(1x2x2)
DoubleConv2D/3D Out_channels = 96
Zoom In
DoubleConv2D/3D Block
Conv2D/3D Kernel=(3x3)/(3x3x3)
GroupNorm Num_groups = 8
Relu
Conv2D/3D Kernel=(3x3)/(3x3x3)
GroupNorm Num_groups = 8
Relu

Figure 3: 2D/3D UNet architecture

Since the focus of this work is leveraging 3D context for MRI segmentation, we use a relatively basic U-Net architecture. Future work could explore more advanced segmentation architectures, such as transformer-based models.

5.3 U-Net Hyperparameters

We performed a grid search on select hyperparameters to determine the optimal configuration for each model type (FLAIR only and pre-contrast+FLAIR+post-contrast). This included the optimizer, learning rate, and batch size. For batch size, we further refined the selection based on chunk size. Other parameters were chosen based on commonly used configurations in the literature.

The hyperparameters of the models are detailed below.

5.3.1 Network Architecture

- Number of Encoder-Decoder Blocks (Depth): 3
- First Encoder Input Channels: 24
- Number of Output Channels per Encoder Block: Doubles at each downsampling step, except for the deepest block, which maintains the same number of channels as the previous one (96).
- Convolutional Layers in Encode/Decoder:
  - 2D Models: convolutional layers with a (3×3) kernel size.
  - 3D Models: convolutional layers with a $(3 \times 3 \times 3)$ kernel size.

5.3.2 Downsampling and Upsampling Layers

- Encoder Downsampling:
  - 2D Models: Max pooling with a (2×2) kernel size.
  - 3D Models: Max pooling with a $(1 \times 2 \times 2)$ kernel size.
- Decoder Upsampling:
  - 2D Models: Bilinear interpolation.
  - 3D Models: Trilinear interpolation.

5.3.3 Normalization and Regularization

- Normalization Layer: Group Normalization (instead of Batch Normalization).
- Number of Groups in Group Normalization: 8.

5.3.4 Training Parameters

- Maximum Training Epochs: 100 (with early stopping).
- Optimizer: Adam.
- Learning Rate:
  - FLAIR-only models: 0.0001
  - Pre-contrast + FLAIR + Post-contrast models: 0.0002

5.3.5 Batch Size

The batch size is dynamically adjusted based on the chunk size to balance memory constraints, computational efficiency, and training stability. The following table presents the batch size configurations:

5.4 Loss Function

From the exploratory data analysis (EDA) in the previous section, we observed a severe class imbalance:

- Tumor-containing slices are significantly fewer than background-only slices.
- Even within tumor-containing slices, most pixels belong to the background class.

To address this class imbalance, we use a composite loss function, combining:

Chunk Size	Batch Size
1	32
3	32
5	32
7	16
9	16
11	16
13	16
15	8
17	8
19	8

Table 1: Batch size configuration for different chunk sizes.

1. Focal loss (to down-weight easy-to-classify background pixels).
2. Dice loss (to improve segmentation performance on minority tumor regions).

The two loss components are equally weighted, and the focal loss parameters are:

- $\alpha = 0.5$
- $\gamma = 2$

5.5 Cross-Validation and Universal Test Set

Our dataset consists of 110 patient scans (one scan per patient), making it relatively small.

To prevent overfitting and data leakage, we first ensure that:

- The dataset is split by patient ID, so that MRI slices from the same patient do not appear in both training and test sets.
- The data is partitioned before preprocessing and chunking, ensuring that correlated slices are not mixed across splits.

5.5.1 Cross-Validation Strategy

Since the dataset is small, we apply K-fold cross-validation to ensure statistically reliable performance estimates while reducing bias.

- We choose K = 5 to balance generalization, reliability, and computational efficiency.
- A universal test set is set aside to evaluate final model performance.

5.5.2 Dataset Splitting

1. 15% of the patients (16 patients) are randomly selected to form the universal test set.
2. The remaining 85% of patients (94 patients) are allocated for training & validation.
3. The 94 training & validation patients are randomly split into 5 non-overlapping folds, i.e. subsets (each containing 18-19 patients).
4. We iterate for each fold:
   - The corresponding subset of the data (19 patients) is used as the validation set.
   - The remaining subsets (75 patients) are used for training.
5. We report mean and standard deviation of results across all folds.

5.6 Ensemble Learning

To further enhance segmentation performance, we apply ensemble learning across models trained with different chunk sizes. Ensemble learning helps improve model performance, increase generalization and reduce overfitting [19].

We explore two ensemble strategies for binary-segmentation:

1. Averaging
   - Pixel-wise averaging of segmentation masks from multiple models and then applying a threshold to determine pixel class.
   - Helps smooth predictions and reduce variance.
2. Voting
   - Hard Voting: Each model predicts a segmentation mask and assigns a class to each pixel after thresholding. The majority class for each pixel is selected. In this work, this is the selected Voting approach.
   - Enhances robustness by prioritizing consensus among models

Both approaches ensure that uncertain predictions are balanced out by more confident ones, leading to more stable and robust segmentations.

6 Experiments

6.1 Evaluation Metrics for Binary Segmentation

To assess the performance of our model, we use several key evaluation metrics that quantify the agreement between predicted and ground truth segmentation masks. The primary metrics include:

Dice Coefficient: The Dice score measures the overlap between the predicted segmentation and the ground truth, defined as:

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where TP (True Positives) represents correctly segmented tumor regions, FP (False Positives) represents incorrectly segmented non-tumor regions, and FN (False Negatives) represents missed tumor regions. The Dice score ranges from 0 to 1, with 1 indicating perfect segmentation.

Intersection over Union (IoU, Jaccard Index): IoU measures the ratio of correctly segmented regions to the total regions present in either the prediction or ground truth:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

IoU is slightly more stringent than Dice, as it does not double-count TP, making it useful for assessing small segmentation errors.

Precision: Precision indicates the proportion of predicted tumor regions that are actually tumors:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

High precision ensures fewer false positives, which is crucial for reducing over-segmentation errors.

Recall: Recall measures the ability of the model to correctly identify all relevant tumor regions:

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

High recall ensures that most of the actual tumor regions are detected, which is particularly important in medical imaging to minimize missed diagnoses.

F1 Score: The F1 score is a harmonic mean of precision and recall, balancing false positives and false negatives:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

The F1 score is useful when dealing with imbalanced data like in our case, as it considers both precision and recall. In the case of binary segmentation, F1 score is equal to Dice score.

By using these metrics, we ensure a comprehensive evaluation of our segmentation model, balancing overlap, false positives, and false negatives to achieve optimal tumor detection.

In our work, we conducted the following experiments:

6.2 The Experiments

6.2.1 FLAIR Only vs. Chunk Size

We evaluate the impact of 3D context when using only the FLAIR images. Different chunk sizes are tested to assess whether incorporating neighboring slices improves segmentation performance compared to a single-slice (2D) approach.

6.2.2 Pre+FLAIR+Post vs. Chunk Size

We extend the analysis to a multimodal setting, incorporating pre-contrast, FLAIR, and post-contrast images. This experiment investigates whether the additional imaging sequences enhance segmentation performance across various chunk sizes.

6.2.3 FLAIR Only Ensemble – Average vs. Voting

We ensemble models trained on select chunk sizes using FLAIR-only input. Two ensembling strategies are compared: (1) averaging predictions across models and (2) majority voting to determine the final segmentation mask.

6.2.4 Pre+FLAIR+Post Ensemble – Average vs. Voting

A multimodal ensemble experiment where models trained with different chunk sizes on pre-contrast, FLAIR, and post-contrast images are combined. We compare averaging predictions against majority voting to determine the most effective ensembling method.

7 Results

7.1 Flair Only validation test:

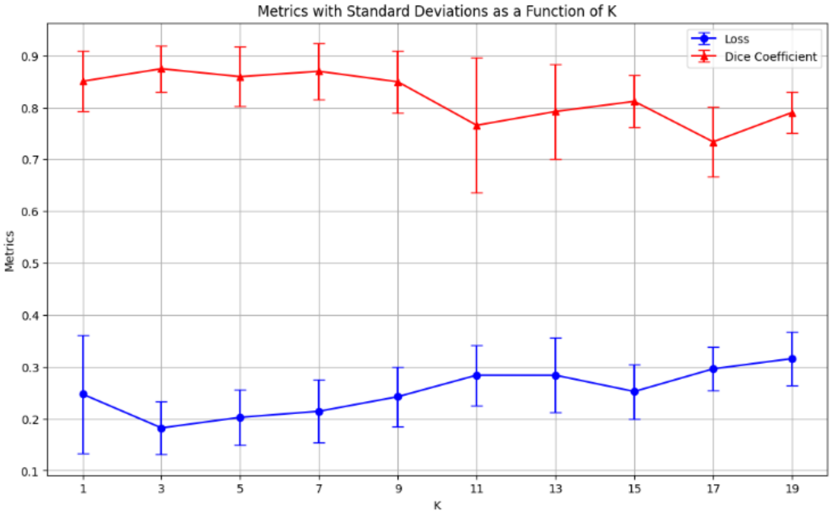

Figure 4: Flair only validation test

7.2 Flair only universal test:

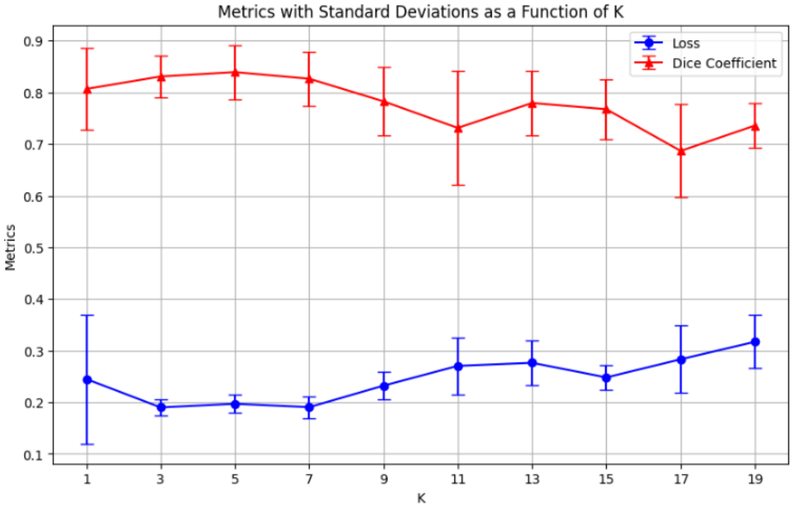

Figure 5: Flair only universal test

From these experiments, we observe that increasing the chunk size initially improves the metrics up to a certain point, after which they begin to decline.

7.3 Pre + Flair + Post validation test:

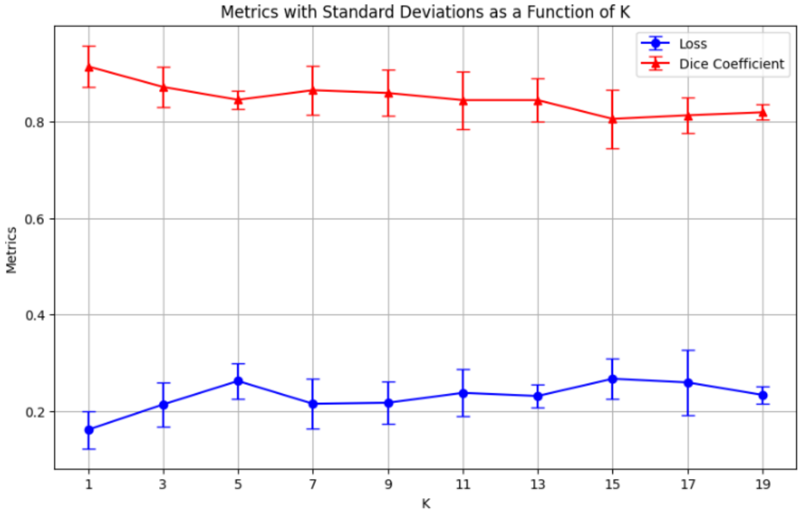

Figure 6: Pre + Flair + Post validation test

7.4 Pre + Flair + Post universal test:

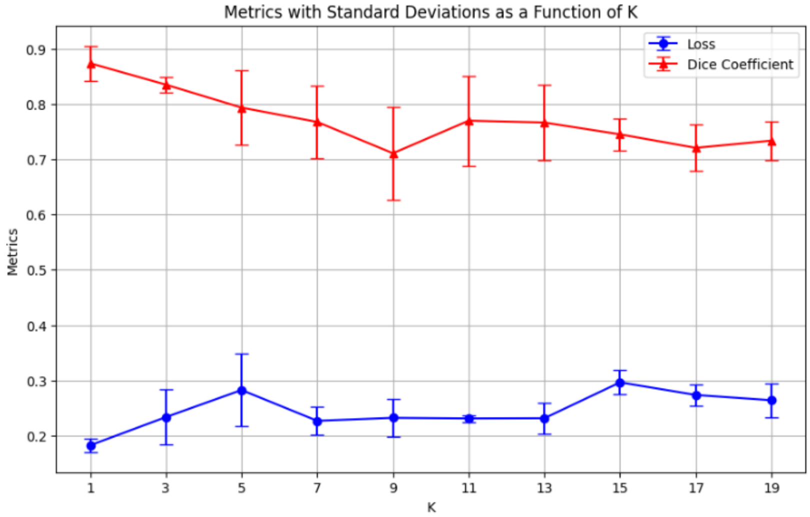

Figure 7: Pre + Flair + Post universal test

From these experiments, we observe that the best results are achieved when the chunk size is 1 (2D). As the chunk size increases, the metrics initially decline until reaching a certain point, after which they relativly stabilize.

7.5 Pre + Flair + Post Vs Flair only (Dice):

As observed, except for the 2D case, the performance of the FLAIR-only models surpasses that of the Pre+FLAIR+Post approach up to and including a chunk size of 9. Beyond this point, the performance becomes inconsistent and lacks a clear trend.

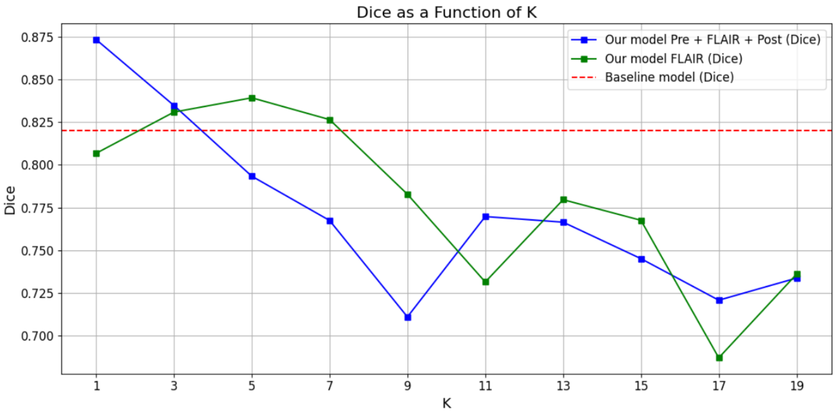

Figure 8: Pre + Flair + Post Vs Flair only Vs Baseline [1] (Dice)

Furthermore, the FLAIR-only model with a chunk size of 1 significantly outperforms the baseline model [1] described in the Related Works section, achieving an improvement of over 5% in Dice score. Additionally, the FLAIR-only models outperform the baseline for chunk sizes between 3 and 7. However, it should be noted that the comparison is not made on the same test set, as the data split in the baseline model was different from ours.

7.6 Flair only average universal heatmap:

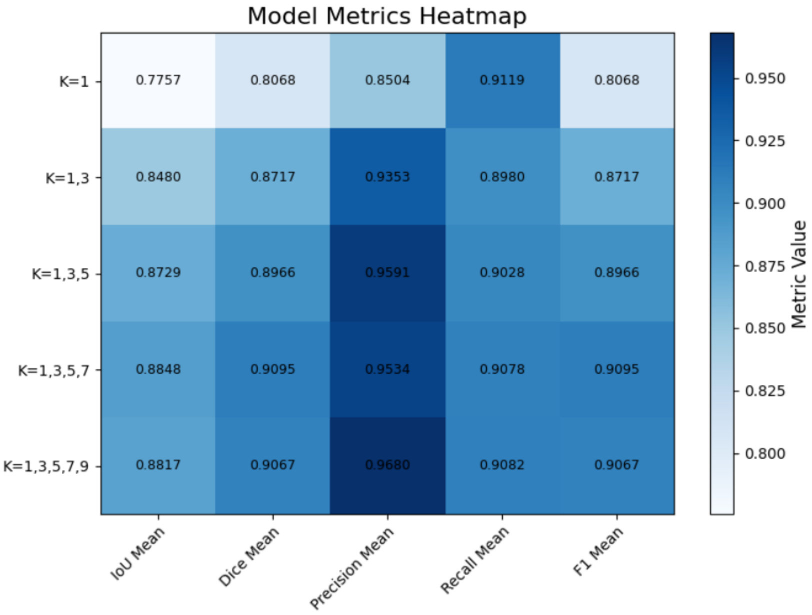

Figure 9: Flair only average universal heatmap

7.7 Flair only voting universal heatmap:

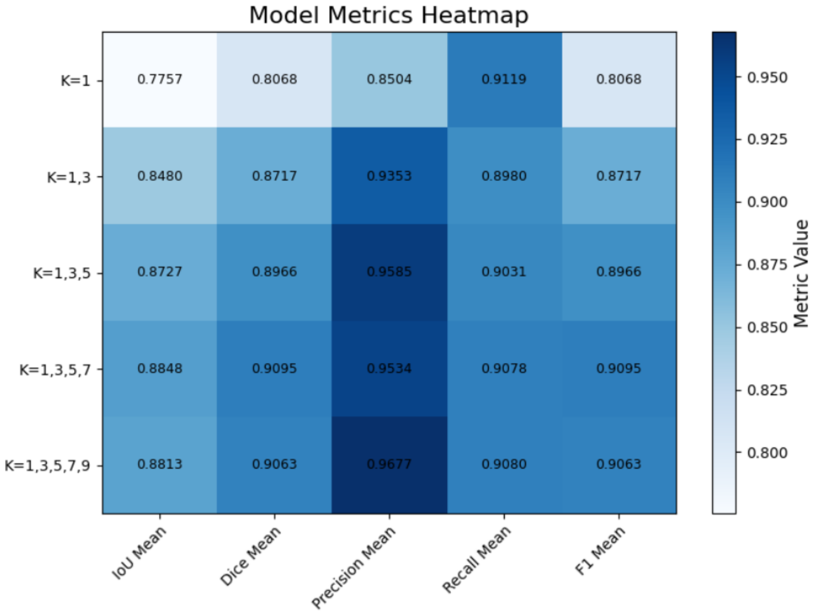

Figure 10: Flair only voting universal heatmap

7.8 Pre + Flair + Post average universal heatmap:

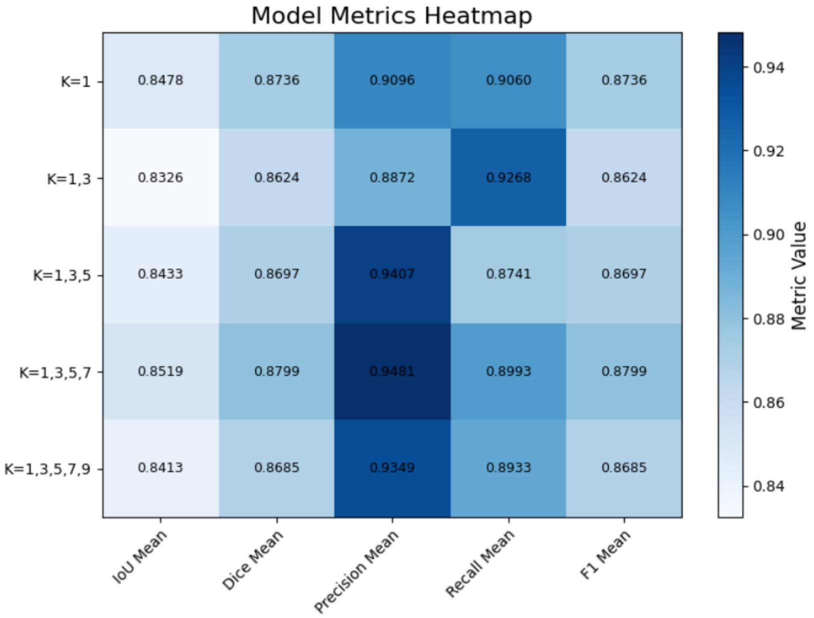

Figure 11: Pre + Flair + Post average universal heatmap

7.9 Model Performance Comparison

Model Type	IoU_{mean}	$\text{Dice}_{\text{mean}}$	$\text{Precision}_{\text{mean}}$	$\text{Recall}_{\text{mean}}$	F1_{mean}
Flair Only Best (K=1,3,5,7)	**0.8848**	**0.9095**	0.9534	0.9078	**0.9095**
Pre+Flair+Post Only Best (K=1,3,5,7)	0.8519	0.8799	**0.9481**	**0.8993**	0.8799
Baseline Model [1]	-	0.8200	-	-	-
Kaggle Model[22]	-	0.8960	-	-	-

Table 2: Performance metrics for different model types. Bold values highlight the best scores for each metric.

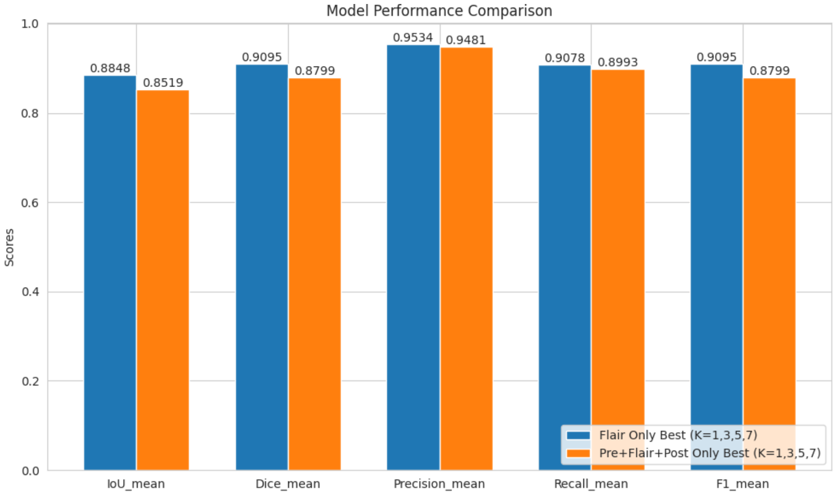

Figure 12: model performance comparison

As we can see from the heatmaps combining models trained on smaller chunk sizes (1, 3, 5, and 7 slices) improves segmentation results for both FLAIR-only and multi-modal models. Additionally, our results indicate that FLAIR-only models outperform multi-modal models when utilizing 3D context alongside ensemble techniques.

Our FLAIR-only Ensemble model significantly outperforms the Baseline model [1], achieving almost a 9% higher Dice score, and also outperforms one of the leading Kaggle notebooks [22] (based on the number of votes) for this dataset by 1.35%. It should be noted that the comparison is not made on the same test set, as the data splits in the baseline and Kaggle models were different from ours.

In the following set of images, we present the predictions of our best model (FLAIR-Only Ensemble). Each image comprises three channels, similar to an RGB image: the pre-contrast agent image in the R channel, FLAIR in the G channel, and the post-contrast agent image in the B channel. Tumor annotations are outlined in red, while the model's predictions are marked in green.

The model accurately detects and segments most tumors. However, some small tumors are not identified correctly, and a few false positives occur, though they are generally near the actual tumor.

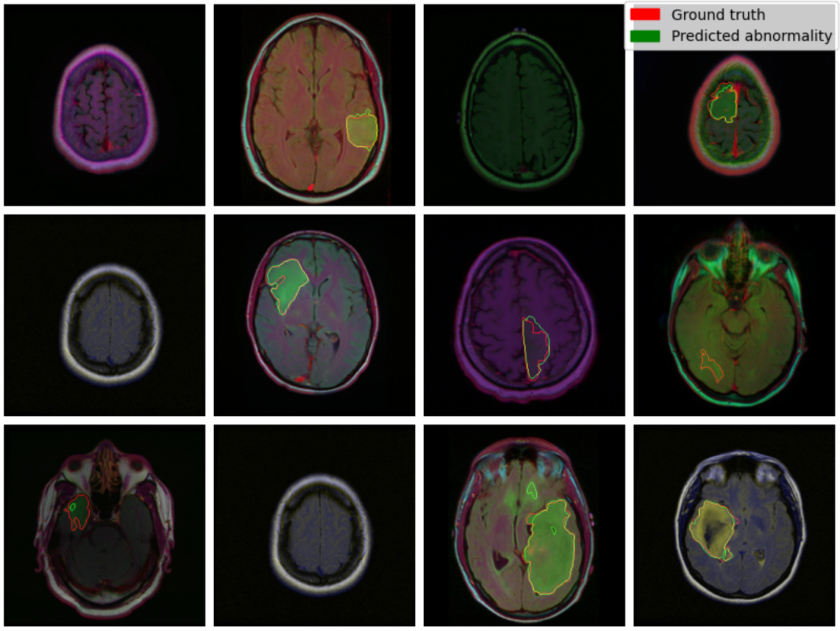

Figure 13: Model Predictions with Tumor Annotations

8 Conclusions

The proposed method offers significant clinical benefits for MRI-based tumor segmentation of Lower Grade Gliomas by achieving accurate segmentation using FLAIR-only MRI, thereby eliminating the need for post-contrast imaging which is commonly used in clinical practice. This could simplify clinical MRI procedures, reduce scanning time, and eliminate the need for contrast agent injection, ultimately minimizing patient discomfort and potential health risks. These advantages make our approach particularly valuable for patients with kidney failure, for whom the injection of a gadolinium-based contrast agent is not recommended.

From a technical perspective, we introduced a method for leveraging 3D contextual information in MRI segmentation by processing input MRI scans in chunks. This approach allows the model to handle varying scan resolutions within the dataset.

Our results demonstrate that for FLAIR-only models, incorporating 3D contextual information improves segmentation performance compared to 2D models when using relatively small chunk sizes. While the improvement in Dice score was noticeable, reaching 3.3%, the results were somewhat surprising, as we anticipated a more substantial gain from utilizing 3D contextual information. However, for larger chunk sizes, the findings aligned more closely with our expectations, showing that the performance gain diminishes as chunk size increases. This is likely because neighboring slices provide relevant spatial information, whereas more distant slices may introduce irrelevant or misleading context.

For multi-modal MRI models incorporating pre-contrast, FLAIR, and post-contrast images, our findings indicate that 2D models outperform all tested 3D models, with segmentation performance decreasing as chunk size increases. This behavior for relatively small chunk sizes differs from that observed in FLAIR-only models, as multi-modal models perform even worse than FLAIR-only models for chunk sizes of 5 to 9 slices. This occurs despite the model having access to the same information as in FLAIR-only models, along with additional information from extra channels (pre-contrast and post-contrast), which could have been ignored to at least achieve similar results as the FLAIR-only models. We believe this results from the model's inability to effectively focus on and learn from informative FLAIR data in neighboring slices when additional MRI channels are present. However, it is important to note that the model's reliance on FLAIR data may have been influenced by the fact that the annotation in the LGG dataset was performed using only FLAIR images. For larger chunk sizes, the behavior of multi-modal MRI models aligns with that of FLAIR-only models, as the spatial relevance of more distant slices diminishes.

To further leverage 3D context, we explored ensemble techniques, including averaging and voting. We found that combining models trained on small chunk sizes (1, 3, 5, and 7 slices) significantly improves segmentation performance for both FLAIR-only and multi-modal models. Notably, our findings indicate that FLAIR-only models outperform multi-modal models when combined with 3D context and ensemble techniques, achieving a Dice score improvement of nearly 3% on average. Additionally, our method significantly outperforms the baseline multi-modal MRI model [1] trained on the same dataset by nearly 9%.

However, given the relatively small dataset used in this study, these findings should be validated on larger datasets to ensure robustness and generalization.

Future research could extend the proposed 3D method, which processes input in chunks, to other medical imaging modalities that produce 3D outputs, such as CT and WSI. Additionally, its application in MRI could be broadened beyond brain tumor segmentation.

Another interesting direction is applying the proposed 3D method in combination with an ensembling technique for models trained exclusively on pre-contrast and post-contrast MRI scans. Furthermore, an ensemble incorporating both types of models could be explored.

An additional improvement could involve incorporating patient metadata (e.g., age, gender, and histologic grade) to enhance segmentation accuracy across patient groups.

9 Code

`https://github.com/shlomi-idc/Final_Project/tree/main`

References

[1] Mateusz Buda, Ashirbani Saha, Maciej A. Mazurowski. "Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm." *arXiv preprint*, Available: `https://arxiv.org/pdf/1906.03720`

[2] Huan Minh Luu, Sung-Hong Park. "Extending nn-UNet for brain tumor segmentation." *arXiv preprint*, Available: `https://arxiv.org/pdf/2112.04653`

[3] Ramy A. Zeineldin, Mohamed E. Karar, Jan Coburger, Christian R. Wirtz, Oliver Burgert. "DeepSeg: Deep Neural Network Framework for Automatic Brain Tumor Segmentation using Magnetic Resonance FLAIR Images." *arXiv preprint*, Available: `https://arxiv.org/pdf/2004.12333`

[4] Fabian Isensee, Paul F. J¨ager, Peter M. Full, Philipp Vollmuth, and Klaus,H.Maier-Hein. "nnU-Net for Brain Tumor Segmentation" *arXiv preprint*, Available: `https://arxiv.org/pdf/2011.00848`

[5] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor K¨ohler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein, "nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation" *arXiv preprint*, Available: `https://arxiv.org/pdf/1809.10486`

[6] Ramy A. Zeineldin, Mohamed E. Karar, Oliver Burgert, Franziska Mathis-Ullrich. "Multimodal CNN Networks for Brain Tumor Segmentation in MRI: A BraTS 2022 Challenge Solution." *arXiv preprint*, Available: `https://arxiv.org/pdf/2212.09310`

[7] Himashi Peiris, Zhaolin Chen, Gary Egan, Mehrtash Harandi. "Reciprocal Adversarial Learning for Brain Tumor Segmentation: A Solution to BraTS Challenge 2021 Segmentation Task." *arXiv preprint*, Available: `https://arxiv.org/pdf/2201.03777`

[8] André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, Jan Egge. "Enhanced Synthetic Data Augmentation and Model Ensemble for Brain Tumor Segmentation." *arXiv preprint*, Available: `https://arxiv.org/pdf/2402.17317`

[9] Muhammad Usmankbar, Måns Larsson, Ida Blystad, Anders Eklund. "Brain tumor segmentation using synthetic MR images – comparison of GANs and diffusion models." *Scientific Data*, Available: `https://www.nature.com/articles/s41597-024-03073-x.pdf`

[10] "LGG MRI Segmentation Dataset." *Kaggle*, Available: `https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation`

[11] "Brain Tumor Segmentation Challenge." Available: `http://braintumorsegmentation.org/`

[12] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, Bo Wang. "Segment Anything in Medical Images." *arXiv preprint*, Available: `https://arxiv.org/pdf/2304.12306`

[13] Jiayuan Zhu,Abdullah Hamdi, Yunli Qi, Yueming Jin, Junde Wu. "Medical SAM 2: Segment Medical Images as Video via Segment Anything Model 2." *arXiv preprint*, Available: `https://arxiv.org/pdf/2408.00874`

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick. "Segment Anything." *arXiv preprint*, Available: `https://arxiv.org/abs/2304.02643`

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv preprint*, Available: `https://arxiv.org/pdf/2010.11929`

[16] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo. "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers." *arXiv preprint*, Available: `https://arxiv.org/abs/2105.15203`

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. "Attention Is All You Need." *arXiv preprint arXiv:1706.03762*, 2017. Available: `https://arxiv.org/pdf/1706.03762`

[18] Olaf Ronneberger, Philipp Fischer, Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation." *arXiv preprint*, Available: `https://arxiv.org/abs/1505.04597`

[19] M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, P.N. Suganthan. "Ensemble deep learning: A review." *arXiv preprint*, Available: `https://arxiv.org/pdf/2104.02395`

[20] Zixin Hao. "Comparative Analysis of Transformer Integration in U-net Networks for Enhanced Medical Image Segmentation." Available: `https://drpress.org/ojs/index.php/HSET/article/view/20596`

[21] `https://seer.cancer.gov/statfacts/html/brain.html`

[22] `https://www.kaggle.com/code/abdallahwagih/brain-tumor-segmentation-unet-dice-coef-89-6`